Linguistic linked open data and under-resourced languages: from collection to application

Steven Moran, Christian Chiarcos

Angaben zur Veröffentlichung / Publication details:

Moran, Steven, and Christian Chiarcos. 2019. “Linguistic linked open data and under-resourced languages: from collection to application.” In *Development of linguistic linked open data resources for collaborative data-intensive research in the language sciences*, edited by Antonio Pareja-Lora, Maria Blume, Barbara Lust, and Christian Chiarcos, 39–68. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/10990.003.0006.

4 Linguistic Linked Open Data and Under-Resourced Languages: From Collection to Application

Steven Moran and Christian Chiarcos

In this chapter, we argue for the adoption and use of Linked Data for linguistic purposes and, in particular, for encoding, sharing, and disseminating under-resourced language data. We provide an overview of linguistic Linked Data in the context of creating datasets of under-resourced languages, and we describe what "under-resourced" language data are, focusing on lexical resources (wordlists and dictionaries) and annotated corpora (glosses and corpora). We discuss aspects of resource integration with two brief case studies of linguistic data sources that have been transformed into Linked Data. Lastly, we describe the state and the bandwidth of applications of Linked Open Data technologies to under-resourced languages in the general context of the Open Linguistics Working Group and the developing Linguistic Linked Open Data (LLOD) ecosystem.

Introduction

Language scientists are increasingly interested in and gleaning the benefits from integration and computing of under-resourced language data. Different users clearly have different data needs; for example, linguists working on typological theory may require broad but not necessarily deep datasets, while computational linguists typically require big data. Regardless, increased access to (interoperable) data is beneficial both for science and for enterprises; in the language resource community, it has been a subject of intense activity over the last three decades, marked by initiatives such as the TEI (since 1987),[1] ISO TC 37/SC 4 (since 2001),[2] the Open Linguistics Working Group (since 2010),[3] as well as several W3C Community and Business groups (the earliest being OntoLex,[4] since 2011).

A more recent trend in this field is the increased adoption of Linked Data for representing language resources, a technology that was originally designed to create synergies between data sources in the Web of Data. Linked Data has been the focus of several workshop series (e.g., Linked Data in Linguistics, annually since 2012; Multilingual Linked Open Data for Enterprises [MLODE], biannually since 2012). At the Ninth International Language Resource and Evaluation Conference (LREC-2014), Linked Data was announced as the hot topic in the language resource community, and, subsequently, it sparked increased activity in workshops, summer schools, and datathons, including the First Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era (CCURL-2014, Reykjavik, Iceland, May 2014), the First Summer Datathon on Linguistic Linked Open Data (SD-LLOD 2015, Madrid, Spain, June 2015), the EUROLAN-2015 summer school on Linguistic Linked Open Data (Sibiu, Romania, July 2015), and the LSA Summer Institute workshop on the Development of Linguistic Linked Open Data (LLOD) Resources for Collaborative Data-Intensive Research in the Language Sciences (LLOD-LSA 2015, Chicago, July 2015).

Because the applications of Linked Data to language resources are manifold (Chiarcos, Nordhoff, and Hellmann 2012), an exhaustive and up-to-date survey is beyond scope for our contribution in this chapter. We thus take a particular focus on an original research problem in linguistics—that is, the investigation of under-resourced languages; we illustrate the potential of Linked Data for statistical approaches in typology and cross-linguistic multivariate methods for investigating worldwide linguistic and cultural diversity.

This involves dealing with the following questions:

- How can collaborative approaches and technologies be fruitfully applied to the development and sharing of resources for under-resourced languages?
- How can small language resources be reused efficiently and effectively, reach larger audiences, and be integrated into applications?
- How can these resources be stored, exposed, and accessed by end users and applications?
- How can research on under-resourced languages benefit from Semantic Web technologies, and specifically the Linked Data framework?

In this chapter, we argue for the benefits of creating and using Linked Data. In particular, Linked Data is a fruitful method for attaining interoperability and creating useful data disseminations of under-resourced languages. Many of these languages are spoken in areas only recently penetrated by technology such as cell phones, and this creates more data and therefore more economic opportunities for people using them.

First, we define what we mean by "under-resourced languages." Then we give a brief, nontechnical introduction to Linked Data and we home in on using Linked Data for linguistic purposes. Next, we provide two short case studies that illustrate the increased opportunity for collaboration when creating under-resourced language data and tools using Linked Data technologies. Later we describe a large in-progress collaborative dataset, the Linguistic Linked Open Data cloud (LLOD), and we introduce the Open Linguistics Working Group (OWLG), a movement led both by computer scientists and linguists aimed at increasing the synergy between research being done in small-scale circles (e.g., field workers and small-scale language documentation projects) and larger and often enterprise-driven initiatives like MLODE or LIDER[5] to support content analytics of unstructured multilingual data. We begin by describing why increased access to under-resourced languages is important. And we end with directions to additional information on Linguistic Linked Open Data, including some do-it-yourself guidelines.

What Are Under-Resourced Languages?

Linguistic Diversity

Even though our view is very far from complete, world-wide linguistic diversity is simply astounding (cf. Evans and Levinson 2009).[6] Given the state of the world's languages, many of which are either endangered or moribund,[7] it is a high priority to document and describe these languages.

With this picture in mind, another fact to bear in mind is the lack of data that would enable us to undertake broad quantitative studies on cross-linguistic diversity. Typologists have coped by using statistical sampling methods to infer characteristics from signals in the genealogical descent or areal contact between languages (Cysouw 2005). This lack of data on the world's languages is referred to as the bibliographic sampling bias. The World Atlas of Language Structures (WALS; Dryer and Haspelmath 2013) is a classic example, at least among typologists, of a convenience sample with over 150 variables, examples being "Word Order" and "Hand and Arm," that necessarily paints an incomplete picture of worldwide linguistic diversity, which in turn spurs qualitative or speculative explanations (McNew, Derungs, and Moran 2018).

The most detailed picture that exists regarding the linguistic documentation of the world's languages is the Glottolog (Nordhoff et al. 2013).[8] Glottolog contains a bibliography about what is currently known about the state of documentation of the world's languages and it is available as Linked Data (Hammarström et al. 2015).[9] But what is known about the documentation of the world's "under-resourced" languages, and how does Linked Data help us combine that data with already existing knowledge?

Under-Resourced Languages

It is clear that languages lacking any documentation whatsoever are "under-resourced," since they are simply *not resourced*, so to speak. There is, however, a notion that there is a set of languages somewhere between very minimally documented ones (say, one grammar or dictionary) and large well-documented languages (examples being Chinese, English, French, German, Russian, and Spanish). This set of languages has been given various labels in the literature. Perhaps the oldest is "low-density languages" (Jones and Havrilla 1998). The terms "medium-density" and "lower-density languages" have also been coined (e.g., Maxwell and Hughes 2006). The latter term specifically refers to "the amount of computational resources available, rather than the number of speakers any given language might have" (Maxwell and Hughes 2006; Meyers et al. 2007). The amount of accessible data, regardless of language-speaker quantities, is the theme that binds these various terms together.[10]

In the language resource community, various categories of "under-resourced" or "weakly supported" languages have been employed:

1. Lack of access to language data—a general lack of language documentation and description (no grammars, dictionaries, or corpora)
2. Lack of access to digital language data—resources exist but cannot easily be accessed
3. Lack of IT/NLP support
4. Limited interoperability of data and tools

For category 1, there are thousands of languages with minimal or no documentation at all. This fact is so clear that we need not list examples.[11]

Category 2 applies to languages for which materials exist but access to those materials is not possible. In the most basic case, there is a lack of access to a digital resource; for instance, some linguist created a corpus of language X using software Y that is now obsolete. Perhaps more often, the case of inaccessibility is due to other factors, such as unsupported character encodings, unavailable fonts, the lack of a standardized orthography, or simply inaccessible data (caused by copyright restrictions, because they are housed in private collections, or only a few paper copies exist, and so on). For audio and video data, the nontransformation from analog to digital (or future) formats, as happened with first reel-to-reel and then cassette tapes, hinders data access.

Category 3 of under-resourced language data is only relevant when the first two points have been addressed. Without localized digital data, language-specific IT/NLP applications cannot exist. In this regard, we see concretely where under-resourced languages lie, as for example the Hausa language which, with some 30 to 50 million speakers, does not possess the digital resources needed for doing basic Natural Language Processing (NLP) tasks.

Category 4 leads us to the final issue in defining under-resourced languages. Technologically, limited interoperability of data and tools is prevalent in many areas, such as tools and annotations, which use different formats and conventions. Until recently, the Russian language has been a prime example; despite being spoken by ~150 million people worldwide, it has until recently lacked large-scale corpora, annotation schemes, and experimental NLP tools . Since the publication of the syntactic annotations of the Russian National Corpus[12] in 2008, the situation is slowly improving. Yet, even the current lack of interoperable digital resources for developing NLP tools exemplifies the point about under-resourced languages raised by Maxwell and Hughes (2006): It is the lack of accessible digital data, not the population of speakers of a given language, that determines whether the language is under-resourced.

Linguistic Resources

Determining under-resourced languages from a computational perspective requires that the resources of a given language be quantified. In this regard, the METANET white papers (Rehm and Uszkoreit 2013) have summarized the status for (most) officially recognized languages in the European Union (EU). The picture is not particularly satisfying. Out of 30 languages, only English is classified as having good support in terms of language resources. In terms of language resources required by different subfields of NLP, half the EU languages have fragmentary support.[13] And only five EU national languages are said to have weak or no support in such resources.[14] Coverage is even more dismal within certain NLP subfields; for example, two-thirds of the languages have weak-to-no support for machine translation. Of course this is the NLP view, where the degree of resource support is estimated from experts' assessment of both the quality/size of digital text, speech, and parallel corpora and their annotations, and of the quality/coverage of machine-readable lexical resources and grammars.

Resource types adopted to define a language as being (under-)resourced in linguistics are somewhat different. Glottolog, as an example, reports on the known language documentation with a focus on grammars, grammar sketches, dictionaries, and wordlists. These resources usually come with qualitative analyses, that is, analyses written by linguists on the basis of certain theoretical preconceptions. By nature, the act of creating a description of a language imposes theoretical constraints on the material collected. In other words, no universally accepted theory exists for describing a language as a system or a model, hence these language resources, even when electronically available, are often not available in a machine-readable format and in any event are usually incompatible with each other. Similar interoperability issues exist between these resources and annotated corpora, with respect to machine-readable dictionaries and grammars required by the METANET definition of "weakly supported" languages.

However, several linguistic data structures have in fact been standardized, to various extents. We focus on lexical resources and annotated (corpus/gloss) data. The third major class of digital language resources—tools for automated and semiautomated annotation—is beyond the scope of this chapter, as it presupposes the availability of dictionaries or corpora.

Lexical Resources: Wordlists and Dictionaries

The wordlist is often considered the most basic linguistic data structure. This generalization is superficial and misses the fact that the wordlist may be more complex than a simple pair of words with labels, such as "gloss" and "word." Yet the question of what a gloss is, is important in defining the nature of the relationship between "gloss" and "word." Perhaps better defined in light of multilingual wordlists is the notion of a "concept" that maps to a particular language-specific form. For example, many languages collapse the notions of "hand" and "arm" (used by English speakers, for example) into one concept that is a single entity. Therefore, there is a mapping relation between certain concepts, as conceptualized in different languages, and their language-specific forms. The relationship between concept and form is neither a definition nor a translation, but rather what has been termed "counterpart" in multilingual comparative contexts (Good 2013).[15]

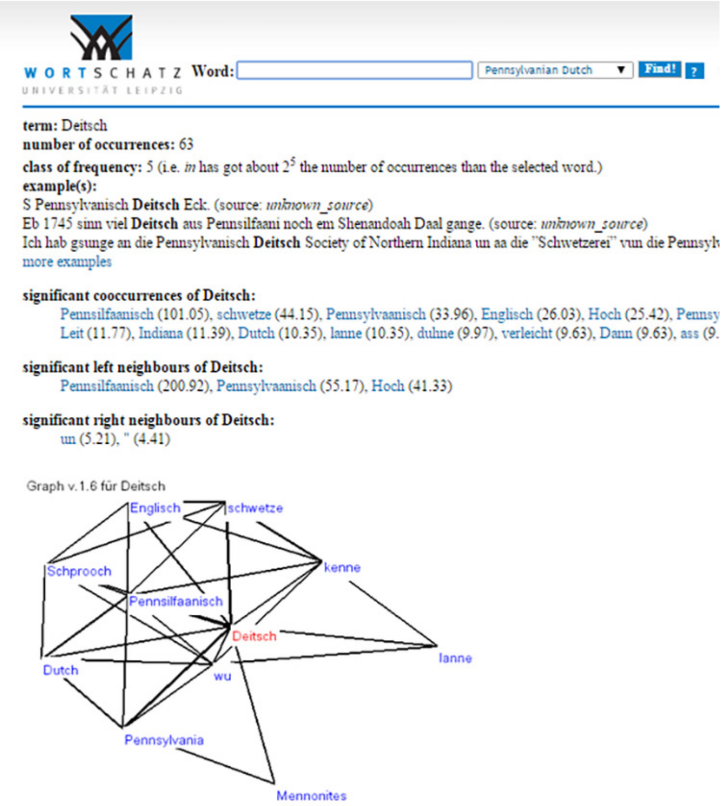
A dictionary is more detailed than a wordlist. It is typically idealized as a collection of form-to-meaning descriptions. Descriptions of forms are typically specified in culturally specific contexts (such as local flora and fauna), which makes it difficult to merge different dictionaries (or lexicons) into one large comparable multilingual source, like a multilanguage wordlist.

For languages that lack manually produced language resources but that come with considerable amounts of digitally available text, another type of lexical resource can be mentioned: frequency and collocation ("association") dictionaries that can be automatically derived from running text (Zock and Bilac 2004). One example is the Wortschatz portal,[16] which provides collocation and frequency dictionaries for 229 languages, including minor languages such as Manx (extinct), Neo-Aramaic (endangered), or Klingon (fictional). Figure 4.1 shows the example entry *Deitsch* "German" from Pennsylvania Dutch (a German dialect spoken in the United States) along with the information provided about it: frequency class (to estimate whether it is has grammatical or lexical function), examples, co-occurring words and frequent collocations, including words of the same semantic class (*Englisch*, *Dutch*, *Schprooch* "language"), related ethnic and geographic concepts (*Pennsylvania, Pennsilfaanisch, Mennonites*), and associated verbs (of speaking, *kenne* "to know," *lanne* "to learn," *schwetze* "to speak"). Although this information does not replace that in a traditional dictionary, it can be used as a tool to construct one, or to confirm the usage of an unknown word (Benson 1990). These resources are also useful for bootstrapping the development of multilingual lexical data translation graphs (cf. Kamholz, Pool, and Colowick 2014).

Annotated Data: Glosses and Corpora

In linguistically annotated data, examples are typically provided in the form of interlinear glossed text (IGT), a semi-standardized data structure comprising three or more lines that prototypically contain three items: an idiosyncratic transcription, a detailed linguistic interpretation (such as a morphological gloss or a part-of-speech tag), and a literal translation.[17] After identification (say, via regular expressions), IGT is automatically extracted from websites and online documents and then assigned an ISO 639-3:2007 language name identifier, derived from attributes identified in the source document. Searching across IGT of thousands of languages in varying detail is desirable, but since the transcription and annotation styles may differ from document to document, some additional layer of what may be called an ontological annotation is needed to logically and consistently define relations in the dataset (cf. Moran 2012a).

Taken a step further, the principle of glossing has been extended to the annotation of larger texts and even entire corpora, as for instance by using tools such as Toolbox.[18] By design, corpora are structured entities consisting of collections of primary data (texts, transcripts, image, audio, or video content), together with their metadata (author, source, date, location, language), and, usually, linguistic annotations as well. Modern corpora have been used as a tool for linguistic research since the Brown Corpus (Kučera and Francis 1967), which has since been compiled as a citation base for the *American Heritage Dictionary*, and which more recently became a cornerstone of corpus linguistics and NLP with the Penn Treebank (Taylor, Marcus, and Santorini 2003) and others.



term: Deitsch
number of occurrences: 63
class of frequency: 5 (i.e. *in* has got about 2^5 the number of occurrences than the selected word.)
example(s):
S Pennsylvanisch **Deitsch** Eck. (source: *unknown_source*)
Eb 1745 sinn viel **Deitsch** aus Pennsilfaani noch em Shenandoah Daal gange. (source: *unknown_source*)
Ich hab gsunge an die Pennsylvanisch **Deitsch** Society of Northern Indiana un aa die "Schwetzerei" vun die Pennsylv
more examples

significant cooccurrences of Deitsch:
Pennsilfaanisch (101.05), schwetze (44.15), Pennsylvaanisch (33.96), Englisch (26.03), Hoch (25.42), Pennsy
Leit (11.77), Indiana (11.39), Dutch (10.35), lanne (10.35), duhne (9.97), verleicht (9.63), Dann (9.63), ass (9.

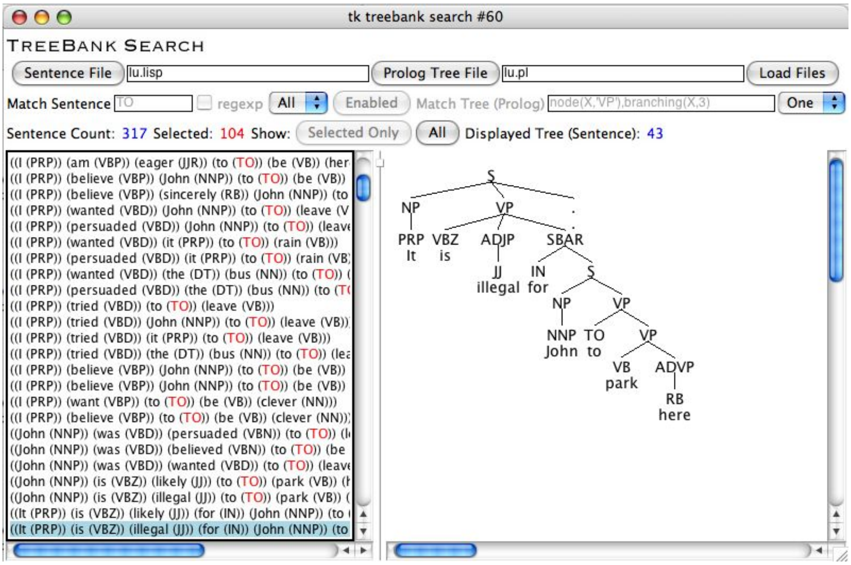
significant left neighbours of Deitsch:
Pennsilfaanisch (200.92), Pennsylvaanisch (55.17), Hoch (41.33)

significant right neighbours of Deitsch:
un (5.21), " (4.41)

Figure 4.1
Example word *Deitsch* ("German") from Pennsylvania Dutch in the Wortschatz portal.

Taking the Penn Treebank as an example, typical annotations comprise lemmatization, morphosyntax (parts of speech, inflectional morphology), syntactic analyses (here phrase structure grammar, otherwise also nominal/clausal chunks or dependency analysis), and, for well-resourced languages, higher levels of analysis such as semantic roles (Kingsbury and Palmer 2002; Meyers, Reeves, Macleod, Szekely, et al. 2004), temporal relations (Pustejovsky et al. 2003), pragmatics (Carlson et al. 2002; Prasad et al. 2008), or co-reference (Pradhan et al. 2007)—in this case specialized subcorpora of the Penn Treebank. Figure 4.2 shows morphosyntactic and syntactic annotations of the Penn Treebank.[19]

For languages without annotated corpora, parallel corpora (such as the Bible, the Qur'an, various translated literature, technical or operational manuals, localization files from software distributions, or subtitles) can be used to bootstrap linguistic annotations via annotation projection (Yarowsky, Ngai, and Wicentowski 2001). Aligned syntactic annotations in a parallel corpus are shown in figure 4.3.[20]



tk treebank search #60

TREEBANK SEARCH

Sentence File lu.lisp Prolog Tree File lu.pl Load Files

Match Sentence TO □ regexp All Enabled Match Tree (Prolog) node(X,'VP'),branching(X,3) One

Sentence Count: 317 Selected: 104 Show: Selected Only All Displayed Tree (Sentence): 43

```
((I (PRP)) (am (VBP)) (eager (JJR)) (to (TO)) (be (VB)) (her
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (believe (VBP)) (sincerely (RB)) (John (NNP)) (to
((I (PRP)) (wanted (VBD)) (John (NNP)) (to (TO)) (leave (V
((I (PRP)) (persuaded (VBD)) (John (NNP)) (to (TO)) (leave
((I (PRP)) (wanted (VBD)) (it (PRP)) (to (TO)) (rain (VB)))
((I (PRP)) (persuaded (VBD)) (it (PRP)) (to (TO)) (rain (VB)
((I (PRP)) (wanted (VBD)) (the (DT)) (bus (NN)) (to (TO)) (
((I (PRP)) (persuaded (VBD)) (the (DT)) (bus (NN)) (to (TO
((I (PRP)) (tried (VBD)) (to (TO)) (leave (VB)))
((I (PRP)) (tried (VBD)) (John (NNP)) (to (TO)) (leave (VB))
((I (PRP)) (tried (VBD)) (it (PRP)) (to (TO)) (leave (VB)))
((I (PRP)) (tried (VBD)) (the (DT)) (bus (NN)) (to (TO)) (lea
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (believe (VBP)) (John (NNP)) (to (TO)) (be (VB))
((I (PRP)) (want (VBP)) (to (TO)) (be (VB)) (clever (NN)))
((I (PRP)) (believe (VBP)) (to (TO)) (be (VB)) (clever (NN))
((John (NNP)) (was (VBD)) (persuaded (VBN)) (to (TO)) (l
((John (NNP)) (was (VBD)) (believed (VBN)) (to (TO)) (be
((John (NNP)) (was (VBD)) (wanted (VBD)) (to (TO)) (leave
((John (NNP)) (is (VBZ)) (likely (JJ)) (to (TO)) (park (VB)) (
((John (NNP)) (is (VBZ)) (illegal (JJ)) (to (TO)) (park (VB)) (
((It (PRP)) (is (VBZ)) (likely (JJ)) (for (IN)) (John (NNP)) (to
((It (PRP)) (is (VBZ)) (illegal (JJ)) (for (IN)) (John (NNP)) (to
```

Figure 4.2
Annotations of the Penn Treebank as visualized by TreeBank Search.

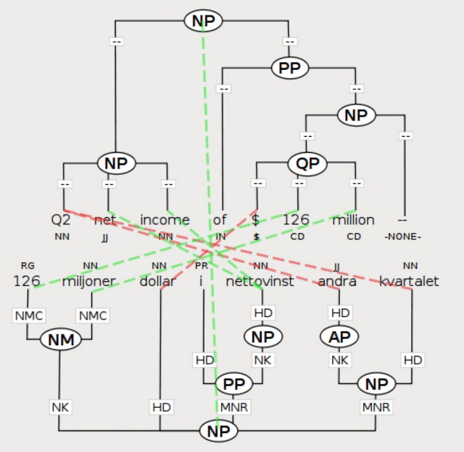


Figure 4.3
Parallel corpus with syntactic annotations and alignment as visualized by TreeAligner.

For languages with a great deal of digitally available text, but lacking NLP support, unsupervised NLP tools may be an option. These extend the concept of collocation extraction to unsupervised grammatical analysis (Clark 2003). However, as this information is only partially interpretable in terms of traditional grammatical categories, and requires considerable amounts of data, this is a current topic of research and beyond the scope of this chapter.

Summarizing, the structures of linguistic resources are manifold even within a single language, and for under-resourced languages resource development even requires links between such structured entities across different languages. Resource integration is thus not only a key problem for modern linguistics in general but also for under-resourced languages in particular.

Resource Integration

It is important to note that linguistic resources are *complex and structured* entities that are composed of different components that need to be integrated if interoperability is to be attained. For example, there is primary data (such as lexemes in a dictionary, text in a corpus, audio or video streams in multimedia corpora), secondary data (including natural language translations, such as glosses and their definitions in a dictionary, or the translation in a parallel corpus or a bilingual wordlist), grammatical analyses (such as in dictionaries, glosses, and annotations), and possibly cross-references (such as a keyword-in-context [KWIC] view in a corpus, a lookup facility from corpus to dictionary to compare the definition of a word, or a lookup facility from dictionary to corpus to provide real-world examples).

Out of this situation of inoperability of data sources and types emerges the challenge to represent (linguistic) data structures on a technical level. Varying solutions to the problem have been proposed, but they have often either been problem-specific (say, a domain-specific [lexicon] XML format via Toolbox) or what might be called “local” (that is, integration within a relational database, showing for instance how to store language and author-specific IGT examples). Each solution probably has its merits; the most widely known solutions have achieved a level of maturity or publicity that has led to their acceptance within their community.

Still, linguistic resources created in an idiosyncratic fashion are not easily reused, unless they can be (easily) integrated with other datasets. This is one of the core functionalities of Linked Data. But at the same time, Linked Data helps us to overcome the heterogeneity of existing formalisms for different local resources, such as dictionaries and corpora. However, existing infrastructures, resources, and tools will continue to be used, and it would be premature to suggest a general shift from existing technology to Linked Data. Instead, we delineate here ways that may be used to automatically convert an existing resource to Linked Data and demonstrate some of the benefits we have gleaned from this conversion.

To summarize, questions of how linguistic data types are transformed into Linked Data are as idiosyncratic as the projects or people who make the design decisions to convert from, say, a linguistic data type A to the Linked Data implementation B. We start with a brief overview of Linked Data and then we show how several datasets have been converted into Linked Data in the Linguistic Linked Open Data (LLOD) cloud.

Linked Data and Under-Resourced Language Data

Linked Data

Linked Data are a set of rules, or "best practices," if you will, for publishing data on the web. Linked Data includes a set of protocols and standards, the purpose of which is to establish links between different datasets. Links are used here broadly; mechanisms provide ubiquitous URI resolution whether a user clicks on a link in his or her browser, or whether computer code automatically crawls through machine interpretable data.

The Linked Open Data paradigm postulates four rules for the publication and representation of web resources:

1. Referred entities should be designated by using URIs.
2. These URIs should be resolvable over HTTP.
3. Data should be represented by means of W3C standards (such as RDF; see below).
4. A resource should include links to other resources.

These rules facilitate information integration, and thus, interoperability, in that they require entities to be addressed in a globally unambiguous way (rule 1 above), that they can be accessed (rule 2) and interpreted (rule 3), and that entities that are associated on a conceptual level are also physically associated with each other (rule 4).

Linked Data is also focused on information integration, and in particular on structural and conceptual interoperability. Linked Data developers strive for **structural interoperability** to attain comparable formats and protocols to access both their own and others' data. A goal is to use the same query language for different datasets, which the user can query across, with or without manipulating the underlying logic (or "semantics") encoded into the (combined) dataset(s) (cf. Moran 2012b).

In the definition of Linked Data, the Resource Description Framework (RDF) receives special attention. RDF was designed to provide metadata about resources that are available either offline (as in books in a library) or online (e-books in a store). RDF provides a generic data model based on labeled directed graphs, which can be serialized in different formats. Information is expressed in terms of *triples*—consisting of a *predicate* (relation, i.e., a labeled edge) that connects a *subject* (i.e., a resource in the form of a labeled node) with its *object* (i.e., another resource or a literal or string). For example, the statement *Christian Chiarcos knows Steven Moran* might be (pseudo)-encoded as a single string consisting of the subject, predicate, and object triple:

```
Subject  http://www.acoli.informatik.uni-frankfurt.de/~chiarcos
Predicate  http://xmlns.com/foaf/0.1/knows
Object  http://www.comparativelinguistics.uzh.ch/de/moran.html
```

As shown, RDF resources (nodes)[21] are represented by *Uniform Resource Identifiers* (URIs), and they are therefore globally unambiguous in the Web of Data (as well as the

“Semantic Web”). Linked Data infrastructure allows resources hosted at different locations to refer to each other, which in turn creates a network of collections of data whose elements are densely interwoven.

Several linearizations for RDF data exist, which differ in readability and compactness. RDF/XML was the original standard for that purpose, but it has been largely replaced by Turtle, a more human-readable format. In Turtle, triples are written as sequences of subject, predicate, and object components, concluded with a final dot.

```
<http://www.acoli.informatik.uni-frankfurt.de/~chiarcos>
<http://xmlns.com/foaf/0.1/knows>
<http://www.comparativelinguistics.uzh.ch/de/moran>.
```

A more compact representation can be achieved using namespace prefixes instead of full URIs:

```
PREFIX acoli: <http://www.acoli.informatik.uni-frankfurt.de/~>
PREFIX cluzh: <http://www.comparativelinguistics.uzh.ch/de/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
acoli:chiarcos foaf:knows cluzh:moran .
```

Several database implementations for RDF data are available, and these can be accessed using **SPARQL** (Prud’hommeaux and Seaborne 2008), a standardized query language for RDF data. SPARQL uses a triple notation similar to Turtle, where properties and RDF resources can be replaced by variables. SPARQL was inspired by Structured Query Language (SQL), in which variables can be introduced in a separate SELECT block, and in which constraints on these variables are expressed in a WHERE block in a triple notation. Thus, for example, we can query for relations between two particular people:

```
SELECT ?relation
WHERE { acoli:chiarcos ?relation cluzh:moran . }
```

SPARQL does not only support running queries against individual RDF databases that are accessible over HTTP (so-called SPARQL endpoints), but it also allows users to combine information from multiple repositories (known as “federation”). RDF can thus be used both to *establish* a network (or cloud) of data collections, and to *query* that network directly.

In this way, Linked Data facilitates the resource accessibility and reusability on different levels (Ide and Pustejovksy 2010):

How to access (read) a resource? (Structural interoperability) Resources use comparable formalisms to represent and to access data (formats, protocols, query languages, etc.), so that they can be accessed in a uniform way and that their information can be integrated with each other.

How to interpret (understand) information from a resource? (Conceptual interoperability) Resources share a common vocabulary, so that linguistic information from one resource can be resolved against information from another resource, e.g., grammatical descriptions can be linked to a terminology repository.

How to integrate (merge) information from different resources? (Federation) Web resources are provided in a way that remote access is supported. Using structurally interoperable representations, a query language with federation support allows the user to run queries against multiple external resources within a single query, and thereby to integrate their information at query time.

In other words, structural interoperability means that resources can be accessed in a uniform way and that their information can be integrated with each other.

Conceptual interoperability is the goal to develop and (re-)use shared vocabularies for equivalent concepts. Shared vocabularies allow the user to run *the same query* across different datasets. Conceptual interoperability, also referred to as semantic interoperability, goes beyond using unified structural data formats and provides a type of label translation with an additional layer of Description Logics, as for example when using OWL-DL to encode datasets.[22]

Again, to make data structurally and conceptually interoperable (to varying degrees), the term *federation* refers to bringing structurally and conceptually interoperable datasets together on the web—publishing data already published on the web, preferably under an open license and with a query interface such as a SPARQL endpoint. Open data is part of the mission of the Open Linguistics Working Group (OWLG), which we describe later in this chapter. First, we highlight the data integration problem and then we discuss Linked Data in the contexts of under-resourced language data and NLP.

Under-Resourced Language Data

The tools used to produce language data and to create and disseminate detailed (and often computationally implemented)[23] linguistic analyses produce a rapidly increasing amount and depth of inoperable datasets. The breadth and depth of ongoing research projects range from many small-scale, single-scientist data collection projects (as in "linguist X works with the last remaining speaker of language Y") to smaller-to-medium-scale corpora collections (say, a one-million-word corpus of X), to larger-to-medium projects that combine many resources (such as CLLD),[24] to large-scale big-data producing efforts (Wiktionary, DBPedia, and the like).

Although the focus of each project differs, all of them gain from more or richer data sources. Among many, notable examples of collections that contain detailed data on under-resourced language data include the ANU Database (Donohue et al. 2013), AUTOTYP (Bickel and Nichols 2015), STEDT (Matisoff 2015), and PHOIBLE (Moran, McCloy, and Wright 2014). A tremendous amount of effort has been put into creating these rich datasets, which are often aimed at collecting linguistic diversity. Each dataset contains sets of languages that are under-resourced, but those data remain in project-specific formats, resulting in insufficient data access, possibilities for sharing, and integration for query and comparison.

Linked Data for Linguistics and NLP

For users wishing to create Linked Data for linguistics, we note that publishing Linked Data allows resources to be globally and uniquely identified such that they can be retrieved through standard web protocols. Moreover, resources can be easily linked to one another in a uniform fashion and thus become structurally interoperable. The five main benefits of Linked Data for linguistics and NLP can be stated as follows (Chiarcos et al. 2013):

Conceptual interoperability: Semantic Web technologies allow users to provide, to maintain, and to share centralized, but freely accessible terminology repositories. Reference to such terminology repositories facilitates conceptual interoperability, since different concepts used in the annotation are backed up by externally provided definitions; these common definitions may be employed for comparison or information integration across heterogeneous resources.

Linking through URIs: URIs provide globally unambiguous identifiers, and if resources are accessible over HTTP it is possible to create resolvable references to URIs. Different resources developed by independent research groups can be connected into a cloud of resources.

Information integration at query runtime (Federation): Along with HTTP-accessible repositories and resolvable URIs, it is possible to combine information from physically separated repositories in a single query at runtime; to wit, resources can be uniquely identified and easily referenced from any other resource on the web through URIs. Similar to hyperlinks in the HTML web, the so-called Web of Data created by these links allows for navigation along these connections, and thereby allows free integration of information from different resources in the cloud.

Dynamic import: When linguistic resources are interlinked by references to resolvable URIs instead of system-defined IDs (or static copies of parts from another resource), one should always provide access to the most recent version of a resource. For instance, for community-maintained terminology repositories like the ISO TC 37/SC 4 Data Category Registry (ISOcat; Windhouwer and Wright 2012; Wright 2004), new categories, definitions, or examples can be introduced occasionally, and this information is available immediately to anyone whose resources refer to ISOcat URIs. To preserve link consistency among Linguistic Linked Open Data (LLOD) resources, however, it is strongly advised to apply a proper versioning system such that backward-compatibility can be preserved: Adding concepts or examples is unproblematic, but when concepts are deleted, renamed, or redefined, a new version should be provided.

Ecosystem: RDF as a data exchange framework is maintained by an interdisciplinary, large, and active community, and it comes with a developed infrastructure that provides APIs, database implementations, technical support, and validators for various RDF-based languages, such as reasoners for OWL. For developers of linguistic resources, this ecosystem can provide technological support or off-the-shelf implementations for common problems; for example, a database can be developed to be capable of supporting flexible, graph-based data structures as necessary for multi-layer corpora (Ide and Suderman 2007).

To these, we may add that the distributed approach of the Linked Data paradigm facilitates the distributed development of a web of resources. It also provides a mechanism for collaboration between researchers who use data, employing shared sets of technologies. One consequence is the emergence of interdisciplinary efforts to create large and interconnected sets of resources in linguistics—and beyond.

These benefits are of particular importance to less-resourced languages. Through recent community efforts such as the OWLG and the emergence of the LLOD cloud, resources from many languages can now be:

- found through central metadata repositories (for the OWLG DataHub),
- accessed by traversing from one resource to another that is linked with it, and
- identified and documented through a set of shared vocabularies

It is important to note at this point that the mere availability of linguistic resources may already improve chances for not just finding but actually *developing* resources for additional under-resourced languages. For example, NLP tools, annotations, and machine-readable lexicons may be *ported* from one language to another, related one. This might not help language isolates, such as Basque or perhaps Etruscan, but it would greatly improve the situation of, say, Faroese if resources from Icelandic can be ported. A similar situation persists for the Bantu languages in Africa, for which a certain degree of NLP support has been achieved only in the nation of South Africa, whereas Bantu languages in most other countries further north have no support at all. In certain respects, these languages are relatively closely related, so that resource porting between languages may be an option.

Examples for such porting approaches include the analysis of Ugaritic (an ancient Semitic language spoken in the second millenium BCE) through resources originally developed for the morphological analysis of Hebrew (Snyder, Barzilay, and Knight 2010) or for approaches to performing character-based translation between related languages, as for example with orthography being "normalized" from a less-resourced language to another; the tool chain developed for the latter case can be applied to the former (Moran 2009; Tiedemann 2012). As a formalism to provide language resources in a structurally and conceptually interoperable way, Linked Data provides a potential cornerstone for future approaches on resource porting across varying languages and domains.

Case Studies

In defining under-resourced languages, we mentioned four key problems: (1) lack of access to language data, (2) lack of access to digital data, (3) lack of IT/NLP support, and (4) limited interoperability of data and tools. We can aim to increase the limited interoperability of data and tools by improving both the conceptual and structural interoperability of existing data sources. This can be undertaken with increased IT/NLP support between languages and projects, which can in turn be used to guide digitization efforts to (partially) compensate for the lack of lexical resources of under-resourced languages.

Efforts to improve conceptual and structural interoperability are exemplified by shared vocabularies; examples include Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010; McCrae, Spohr, and Cimiano 2011; lexicons), Lexvo[25] (de Melo 2015) and Glottolog[26] (Hammarström et al. 2015; language identification), PHOIBLE Online[27] (Moran, McCloy, and Wright 2014; phonemes), and OLiA (Chiarcos 2008; annotations). Other efforts to increase the lack of lexical resources are exemplified by projects like QuantHistLing (see below), PanLex[28] (Kamholz, Pool, and Colowick 2014), and LiODi.[29] In this section we provide examples in the form of brief case studies.

QuantHistLing

Projects like QuantHistLing (Quantitative Historical Linguistics)[30] illustrate the effort needed to make linguistically diverse samples of lexical data available to a broad and computationally savvy audience. Any project must first identify the linguistic data sources (such as wordlists and dictionaries) that it wishes to use or to create. QuantHistLing has digitized about 200 source documents, most of them available only in print and many of them the sole resources available for the poorly described and under-resourced languages that they describe. Two examples, one of a comparative wordlist and the other of a bilingual dictionary, respectively, are shown in figure 4.4.

The digitization pipeline involves transforming printed sources into electronic sources (whether by OCR or by manual typing). Once sources exist in an electronic form, for dictionaries the interesting parts of each entry are identified, typically with source-specific regular expressions, to extract head words, translations, example sentences, and part-of-speech information. For wordlists, concepts and their glosses are extracted. Standoff annotations may be added to the data by project members; for example, the “dictinterpretation” data type is added by project members and may include manual corrections or other pertinent information.

The QuantHistLing project produces a simple data output format that contains metadata (prefixed with the symbol “@”) and tab-delimited lexical output on a source-by-source basis.[31] An example is given in figure 4.5.

Using the comma-separated values (CSV) data as input, a simple script was written to transform the data into RDF. An RDF model that is specified in the Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010; McCrae, Spohr, and Cimiano 2011) was created for the QuantHistLing data (Moran and Brümmer 2013). *Lemon* is an ontological model for modeling lexicons and machine-readable dictionaries for linking to both the Semantic Web and the Linked Data cloud. The QuantHistLing-lemon model is illustrated in figure 4.6.

Given the goals of QuantHistLing to uncover and clarify phylogenetic relationships between languages, the transformation of wordlist data and of dictionary data from numerous source documents to an RDF graph provides researchers with a structurally

Chocó

DR	hĩɾṹ
CT	hḛ̃ɾṹ
CM	hĩ́ɾũ
TD	hĩ́ɾã, ɓɨ́ɾɨ
EP	hĩ́ɾũ
BA	bɨ́ɾɨ ekʰára
WM	bɨ

Chibcha

IK	kə́ttɨ
KO	kása
DM	kɨsá
CL	kássa
TN	kes-kára
BI	kixturə

Barbacoa

PA	tʃida
GU	katsik
TR	kaˈtsik
AW	mittɨ
TP	nede
CH	neepa

Kamsá

KS	ʃekuá-tçe

Quechua

IN	tʃáki

Arawak

WY	wóˀui (wa-óˀui)
AC	-íiba
CR	no-iipa
PP	wàabàli (wa-àbàli)
YC	weˀemá (wa-iˀimá)
TO	pititáβe, pititáwe†
CA	hiipa
BN	-ipa
RE	-hiiˀpú

Tucano

TC	dɨˀpó-kã
WN	daˀˈpo-ro
PY	daˀˈpokã
WA	dɨˈpó
BR	dɨˈpo
TY	dɨˈpó
YR	ˈdɨ́po
DE	ˈgúbú-ru
SR	guˀˈbú
TA	rɨˈpó
CP	rɨˈpó
MA	gɨbo
BS	gɨbó
TM	ũˀˈpu-a
CU	kɨˈbó-ba
KG	ˈkũˀa-pɨ
SI	ˈgɨ̃õ-bɨ
SE	ˈkɨ̃õhawa
OR	ɨ̃õ-pɨ

Carib

CJ	ˈhuhu
YK	úʃi

Guahibo

PL	pe-táxu
GH	pe-táxu
CI	pe-táxu
JT	pe-tkút
GY	peh tɨ́ak

Sáliba-Piaroa

SL	haˀba

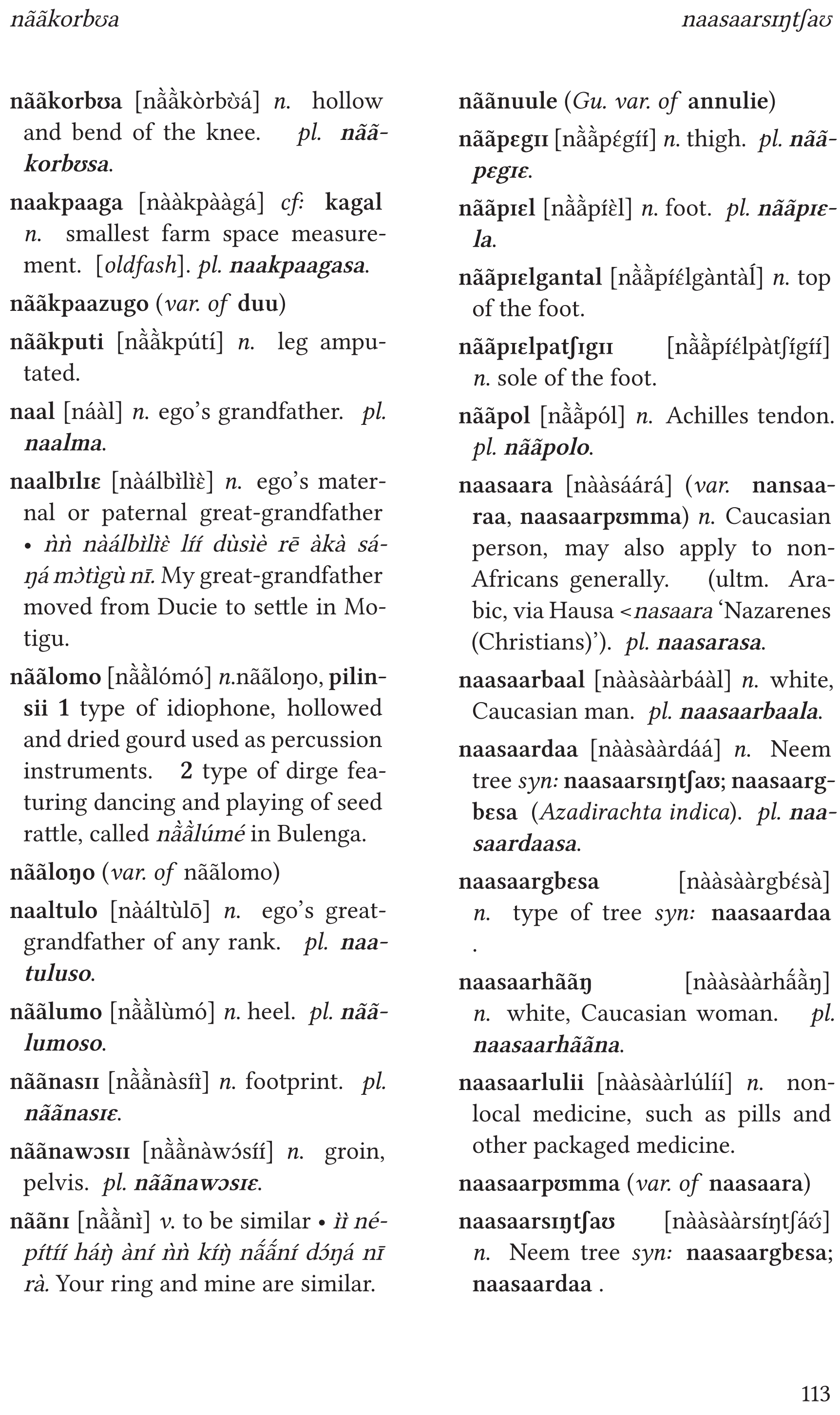
Macú-Puinave

PU	sim
NK	tʃɨ̃i^{4}at^{1}
KK	hit^{2}-tʃa^{4} daˀ^{4}
JU	tʃib

Witoto

MR	e.ɯ-ʤɯ
MN	é.ɯba
NP	e.ɯ́-ba
OC	ɯˀjóó(ga)
MU	tɨ́-ˀai
BO	(mé)-xtʰɯ́ˀaá
MÑ	tʰɯ́ˀaá, íhtʰʲɯ́ˀa

TO Giacone: ue-hépama



nããkorbʊa *naasaarsɪŋtʃaʊ*

nããkorbʊa [nȁȁkòrbʊ̀á] *n.* hollow and bend of the knee. *pl.* ***nããkorbʊsa***.

naakpaaga [nààkpààgá] *cf:* **kagal** *n.* smallest farm space measurement. [*oldfash*]. *pl.* ***naakpaagasa***.

nããkpaazugo (*var. of* **duu**)

nããkputi [nȁȁkpútí] *n.* leg amputated.

naal [náàl] *n.* ego's grandfather. *pl.* ***naalma***.

naalbɪlɪɛ [nàálbɪ̀lɪ̀ɛ̀] *n.* ego's maternal or paternal great-grandfather • *ǹǹ nàálbɪ̀lɪ̀ɛ̀ líí dùsìè rē àkà sáŋá mɔ̀tìgù nī.* My great-grandfather moved from Ducie to settle in Motigu.

nããlomo [nȁȁlómó] *n.*nããloŋo, **pilinsii** **1** type of idiophone, hollowed and dried gourd used as percussion instruments. **2** type of dirge featuring dancing and playing of seed rattle, called *nȁȁlúmé* in Bulenga.

nããloŋo (*var. of* nããlomo)

naaltulo [nàáltùlō] *n.* ego's great-grandfather of any rank. *pl.* ***naatuluso***.

nããlumo [nȁȁlùmó] *n.* heel. *pl.* ***nããlumoso***.

nããnasɪɪ [nȁȁnàsíì] *n.* footprint. *pl.* ***nããnasɪɛ***.

nããnawɔsɪɪ [nȁȁnàwɔ́síí] *n.* groin, pelvis. *pl.* ***nããnawɔsɪɛ***.

nããnɪ [nȁȁnì] *v.* to be similar • *ɪ̀ɪ̀ népítíí háŋ̀ àní ǹǹ kíŋ̀ nȁȁní dɔ́ŋá nī rà.* Your ring and mine are similar.

nããnuule (*Gu. var. of* **annulie**)

nããpɛgɪɪ [nȁȁpɛ́gíí] *n.* thigh. *pl.* ***nããpɛgɪɛ***.

nããpɪɛl [nȁȁpíɛ̀l] *n.* foot. *pl.* ***nããpɪɛla***.

nããpɪɛlgantal [nȁȁpíɛ́lgàntàl̂] *n.* top of the foot.

nããpɪɛlpatʃɪgɪɪ [nȁȁpíɛ́lpàtʃígíí] *n.* sole of the foot.

nããpol [nȁȁpól] *n.* Achilles tendon. *pl.* ***nããpolo***.

naasaara [nààsáárá] (*var.* **nansaaraa**, **naasaarpʊmma**) *n.* Caucasian person, may also apply to non-Africans generally. (ultm. Arabic, via Hausa <*nasaara* 'Nazarenes (Christians)'). *pl.* ***naasarasa***.

naasaarbaal [nààsààrbáàl] *n.* white, Caucasian man. *pl.* ***naasaarbaala***.

naasaardaa [nààsààrdáá] *n.* Neem tree *syn:* **naasaarsɪŋtʃaʊ**; **naasaargbɛsa** (*Azadirachta indica*). *pl.* ***naasaardaasa***.

naasaargbɛsa [nààsààrgbɛ́sà] *n.* type of tree *syn:* **naasaardaa** .

naasaarhããŋ [nààsààrhȁȁŋ] *n.* white, Caucasian woman. *pl.* ***naasaarhããna***.

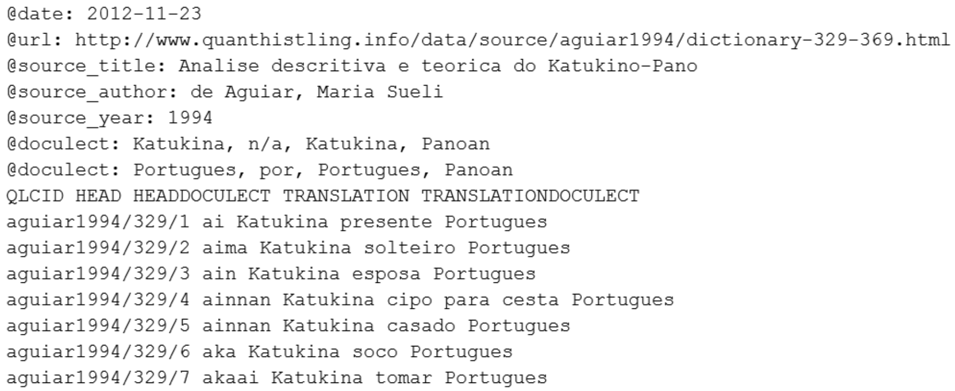
naasaarlulii [nààsààrlúlíí] *n.* non-local medicine, such as pills and other packaged medicine.

naasaarpʊmma (*var. of* **naasaara**)

naasaarsɪŋtʃaʊ [nààsààrsíŋtʃáʊ́] *n.* Neem tree *syn:* **naasaargbɛsa**; **naasaardaa** .

113

Figure 4.4
Wordlist and dictionary exemplars (above and opposite).



```
@date: 2012-11-23
@url: http://www.quanthistling.info/data/source/aguiar1994/dictionary-329-369.html
@source_title: Analise descritiva e teorica do Katukino-Pano
@source_author: de Aguiar, Maria Sueli
@source_year: 1994
@doculect: Katukina, n/a, Katukina, Panoan
@doculect: Portugues, por, Portugues, Panoan
QLCID HEAD HEADDOCULECT TRANSLATION TRANSLATIONDOCULECT
aguiar1994/329/1 ai Katukina presente Portugues
aguiar1994/329/2 aima Katukina solteiro Portugues
aguiar1994/329/3 ain Katukina esposa Portugues
aguiar1994/329/4 ainnan Katukina cipo para cesta Portugues
aguiar1994/329/5 ainnan Katukina casado Portugues
aguiar1994/329/6 aka Katukina soco Portugues
aguiar1994/329/7 akaai Katukina tomar Portugues
```

Figure 4.5
QuantHistLing data extraction format.

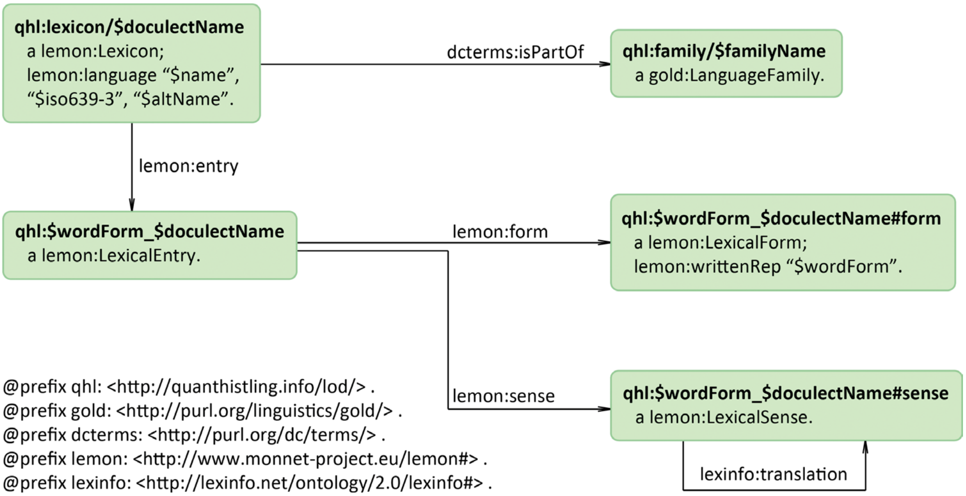

qhl:lexicon/$doculectName
a lemon:Lexicon;
lemon:language "$name",
"$iso639-3", "$altName".
dcterms:isPartOf
qhl:family/$familyName
a gold:LanguageFamily.
lemon:entry
qhl:$wordForm_$doculectName
a lemon:LexicalEntry.
lemon:form
qhl:$wordForm_$doculectName#form
a lemon:LexicalForm;
lemon:writtenRep "$wordForm".
lemon:sense
qhl:$wordForm_$doculectName#sense
a lemon:LexicalSense.
lexinfo:translation
@prefix qhl: <http://quanthistling.info/lod/> .
@prefix gold: <http://purl.org/linguistics/gold/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix lemon: <http://www.monnet-project.eu/lemon#> .
@prefix lexinfo: <http://lexinfo.net/ontology/2.0/lexinfo#> .

Figure 4.6
An implementation of QuantHistLing data modeled in *lemon*.

interoperable resource that we call a *translation graph*—an RDF model that allows users to query across the underlying lexicons and dictionaries to extract semantically aligned wordlists via their glosses and translations.[32] Identifying semantically related sets of words from different languages is one step in investigating the historical evolution of languages and their possible relatedness.[33]

Conversion of wordlist and dictionary data from QuantHistLing into *lemon* has the advantage that *lemon* is tightly integrated with Semantic Web technologies. In particular, lexical data in *lemon* are easily made interoperable with the Linguistic Linked Open Data

(LLOD) cloud. Thus, the resulting lexical resource is available on the web in a standard format and accessible, the data can be made query-able via a SPARQL endpoint,[34] and the use of the *lemon* ontology with Linked Data assists QuantHistLing in its goals to merge disparate dictionary and wordlist data via semantic sense and meaning mappings into an ontology for graph-to-CSV extraction of multilingual and disparate resources.[35]

This is indicative of researchers' efforts at transforming multilingual lexical datasets into Semantic Web data. That is, there exists some input data format (often CSV) from which lexical semantic data needs to be mapped to similar nodes in a given translation graph. Furthermore, metadata about languages or resources in the dataset must be annotated with URIs so that those resources can be linked to other datasets. This linking lies at the heart of the Linked Data initiative, and in particular of the LLOD, which aims to make available an increasing number of resources on under-resourced languages to research communities via the web.

PHOIBLE in CLLD

The PHOIBLE database is a broad collection of spoken languages' phonological systems.[36] It encodes a theory of linguistic description that includes systems of phonemes, allophones, and their phonological conditioning environments. The formalism is known as distinctive feature theory, is semi-binary, and has been used to model broad-base applications for automatic spoken-language (even dialect) recognition. Distinctive feature theory in phonology was developed in the early-to-mid-20th century as an abstraction of the physical acoustic signals (in speech) into a graphemic-based encoding (that is, letter-based transcription) of sounds and their contrasts. This theory allows linguists to describe and predict (un)natural classes of sound changes.

PHOIBLE was initially published as Linked Data in a simple RDF model, which includes concepts (languages, sounds, and features) and the relationsbetween languages and their sounds, and sounds and their features (Moran 2012a, 2012b). This prototype was created by scripting input in CSV data and outputting an RDF graph, given a model, into an XML serialization. More recently, the PHOIBLE data has been incorporated into the Cross-Linguistic Linked Data (CLLD) framework (Forkel 2014). For under-resourced languages, the CLLD framework provides several straightforward mechanisms for taking structured data (say, CSV and BibTeX for bibliographic references), especially from diverse linguistics datasets like typological databases,[37] and generating end-user-friendly interfaces with features like explorable maps, sortable features, and searchable content.[38]

Beyond just a nice web interface, CLLD applications provide their data as Linked Data described with VoID descriptions, and those data are accessible through tools such as rdflib[39] and Python.[40] The core CLLD data model is illustrated in figure 4.7, which contains concepts (Dataset, Language, Parameter, ValueSet, Value, Unit, UnitParameter, UnitValue, Source, Sentence, Contribution) and the relations between entities—providing a triples model (Forkel 2014).[41]

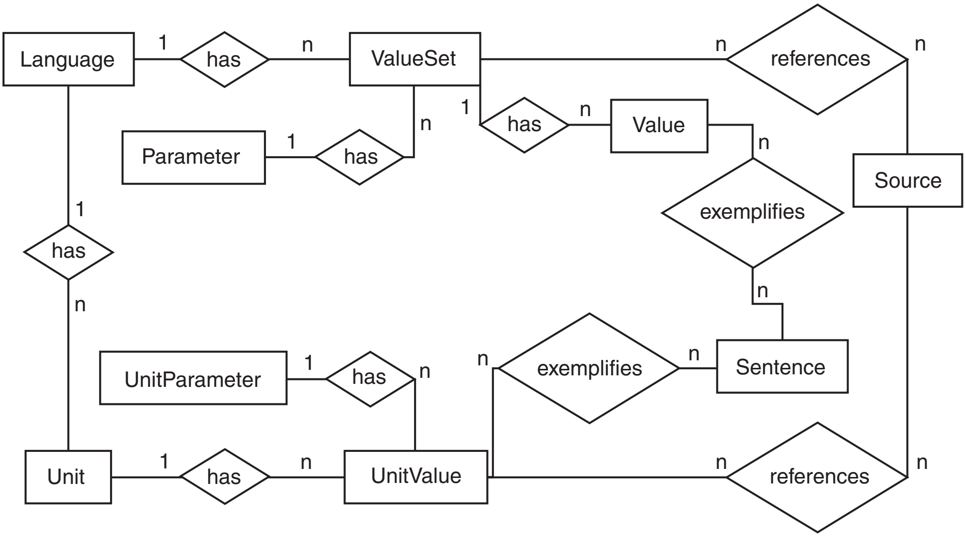


Figure 4.7
Entity-relationship diagram of the CLLD core data model.

The impact of CLLD applications is spelled out in Forkel (2014). In sum, queries like "give me all information on language X" are possible, and they will return all information from all CLLD applications for a given language. The query functionality also allows for testing conjectures made in particular sources, such as the WALS chapter "Hand and Arm" (Brown 2013), on the evolution of languages and other aspects of linguistic diversity. More complex queries that federate the CLLD resources are also possible via the CLLD Portal.[42] Extracted data can then be used either to seed or to expand the development of other datasets with language metadata, linguistic features, and lexical and orthographic encoded data—in particular, data on under-resourced languages that may be used in social media outlets such as social networks, blogs, or tweets.

Combining Case Studies

We have already presented two brief case studies of the transformation of linguistic data into Linked Data. Now we may ask, what can we do with these resulting Linked Data resources? One idea is that we might want to reconsider the notion of resource porting through character-based machine translation. For example, using the PHOIBLE vocabulary, we can describe languages on the level of their phonemic structure and, subsequently, we can also describe the systematic sound correspondences between different languages. We have an appropriate target dataset in QuantHistLing.

At the moment, character-based machine translation manages to identify corresponding characters or character groups, yet treats them as opaque signs. In fact, however, sound correspondences tend to reflect systematic laws, meaning that not one specific phoneme developed into another, but that *all phonemes* with a specific feature turned into phonemes whose feature value was replaced by another value. Unlike state-of-the-art character-based models, a phoneme-level model would be able to capture this information if a mapping from character to phoneme (or phonetic feature set) can be established.[43] This is, however, a direction for future research, and it requires a close integration of linguistic and NLP expertise. Under the umbrella of the interdisciplinary Open Linguistics Working Group (OWLG), however, such a collaboration may be possible, because it represents one of the very few forums where both communities actually meet.

The Linguistic Linked Open Data Cloud

Recent years have seen not only a number of approaches to provide linguistic data as Linked Data, but also the emergence of larger initiatives that aim at interconnecting these resources. Among these, the Open Linguistics Working Group (OWLG) of the Open Knowledge Foundation (OKFN) has spearheaded the creation of new data and the republishing of existing linguistic resources as part of the emerging Linguistic Linked Open Data (LLOD) cloud. These initiatives provide technological infrastructure and community support for researchers wishing to produce and share under-resourced language data.

The LLOD Cloud

Aside from benefits arising from the actual *linking* of linguistic resources, various linguistic resources from very different fields have been provided in RDF and related standards over the last decade. In particular, this is the case for lexical resources like WordNet (Gangemi, Navigli, and Velardi 2003), which represents a cornerstone of the Semantic Web and is firmly integrated in the Linked Open Data (LOD) cloud. In a broader sense, LOD general knowledge bases from the LOD such as the DBpedia have also been rendered as lexical resources, owing to their immanent relevance for Natural Language Processing tasks such as Named Entity Recognition (NER) or Anaphora Resolution (AR). Other types of linguistically relevant resources with less importance to AI and knowledge representation, however, are not a traditional part of the LOD cloud, although they do motivate the creation of a sub-cloud dedicated to linguistic resources.

Figure 4.8 illustrates the Linguistic Linked Open Data (LLOD) cloud diagram. The LLOD cloud is a collection of linguistic resources that are published (typically) under open licenses as Linked Data. The data are decentralized, developed, and maintained with metadata online.[44] The cloud diagram is developed as a community effort in the context of OWLG and is built automatically from metadata about Linked Data sources stored online. Users who wish to have their datasets included need to make sure that at least one URL provided for data or endpoints is up and running. Metadata tags for discoverability include

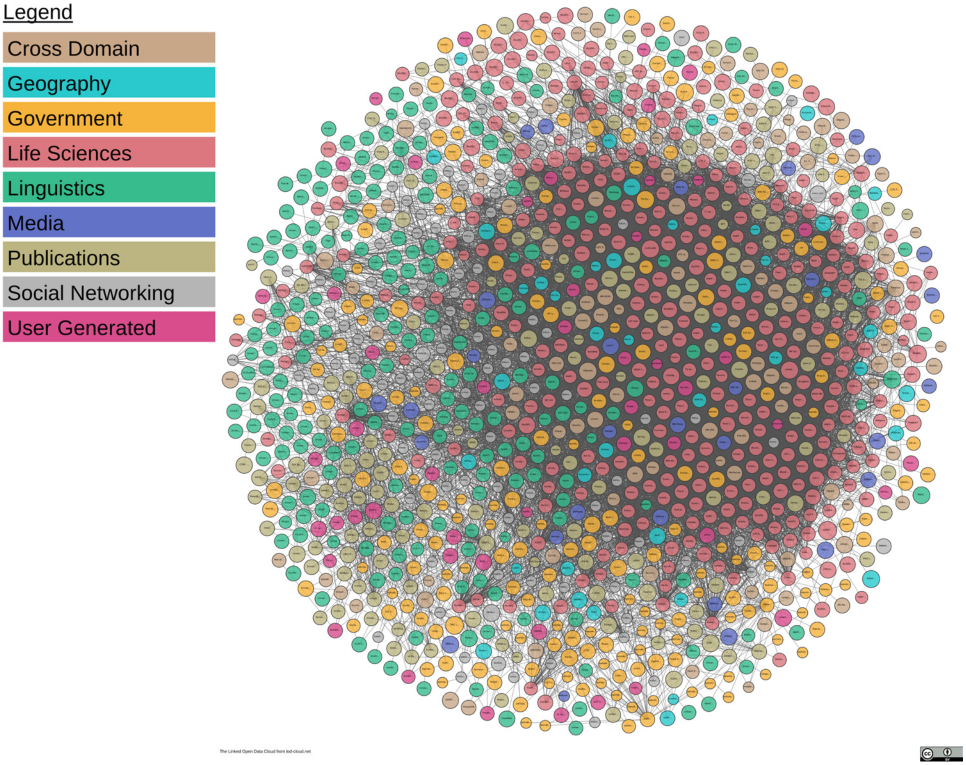


Figure 4.8
Linguistic Linked Open Data (LLOD) cloud.

"llod" and "linguistics." Other tags are used to more precisely define specific resources (e.g., corpus, lexicon, wordnet, thesaurus).

The Open Linguistics Working Group

The LLOD cloud is a result of a coordinated effort by the Open Linguistics Working Group (OWLG; see Chiarcos and Pareja-Lora, this volume).

Since its formation in 2010, the OWLG has grown steadily. One of our primary goals is to attain openness in linguistics through:

1. Promoting the idea of open linguistic resources
2. Developing the means for the representation of Open Data
3. Encouraging the exchange of ideas across different disciplines

Publishing linguistic data under open licenses is an important issue in academic research, as well as in the development of applications. We see increasing support for this in the linguistics community (Pederson 2008), and there are a growing number of resources published under open licenses (Meyers et al. 2007). Publishing resources under open licenses offers many advantages: For instance, freely available data can be more easily reused, double investments can be avoided, and results can be replicated. Also, other researchers can build on the data and subsequently can refer to the publications associated with them. Nevertheless, a number of ethical, legal, and sociological problems are associated with Open Data,[45] and the technologies that establish interoperability (and thus reusability) of linguistic resources are still under development.

The OWLG represents an open forum for interested individuals to address these and related issues. At the time of writing, the group consists of about 100 people from 20 different countries. Our group is relatively small, but continuously growing and sufficiently heterogeneous. It includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology; the ground for fruitful interdisciplinary discussions has been laid out. One concrete result emerging out of collaborations between a large number of OWLG members is the LLOD cloud, as already sketched above. Independent research activities of many community members involve the application of RDF/OWL to represent linguistic corpora, lexical-semantic resources, terminology repositories, and metadata collections about linguistic data collections and publications. To many such members, the Linked Open Data paradigm represents a particularly appealing set of technologies. Within the OWLG, these activities have converged toward building the cloud.

Under-Resourced Languages in the LLOD Cloud

Two principal driving forces of the growth of the LLOD cloud diagram and the OWLG have been, first, the synergies between independent research projects whose experts were interested in providing their data as RDF or Open Data, and, second, multinational projects, often funded by the EU, that focus on technological solutions for multilinguality issues in the European digital single market (affecting matters of localization, computational lexicography, and machine translation). A third factor that contributed to this development has been more recent projects and applications in the humanities and academic branches of linguistics. With the research described in this paper, we demonstrate the applicability of LLOD technologies to one of these “small” areas of research and their ability to harness their highly specific resources in studying under-resourced languages. We consider the adaptation of this technology in an area where both experts and students are often lacking programming skills to be a particularly strong case for the potential of Linked Data in linguistics.

However, the QuantHistLing projects and CLLD are only two exemplary case studies from this particular area. Related efforts that employ RDF and/or Linguistic Linked Open Data for the study and comparison of less-resourced languages include, for example, the “Typology Tool” TYTO (Schalley 2012) that utilizes Semantic Web technologies to process, integrate, and query cross-linguistic data. The Typological Database System[46] (Dimitriadis et al. 2009) uses OWL ontologies for harmonizing and providing access to distributed databases that are created in the course of typological research and language documentation. For a similar application in language resource harmonization, the GOLD ontology was created as part of the Electronic Metastructure for Endangered Languages Data (E-MELD, see Langendoen, this volume).

Poornima and Good (2010) have already described the application of RDF and Linked Data technologies for creating machine-readable wordlists for under-resourced languages. Building on these and other pieces of earlier research, the project called Linked Open Dictionaries (LiODi) is currently developing techniques to facilitate cross-linguistic search across dictionaries to assist in language contact studies among endangered and historical languages in the Caucasus area and among Turkic languages (Abromeit et al. 2016), as well as to assist in the LLOD conversion of formats typically used in linguistic typology and for language documentation (Chiarcos et al. 2017). While these technologies and the resources created on this basis are still under development, the PanLex project (Kamholz, Pool, and Colowick 2014) has already published a near-universal RDF-based translation graph that covers numerous under-resourced languages.

Getting Additional Guidance

As is the case when experts adopt any state-of-the-art technologies, advances and developments are happening faster than traditional print media can possibly keep up with. In this paper, we provided sound reasoning and examples of why we believe Linked Data is an important platform for working with and disseminating under-resourced language data. Nevertheless, the tools and technologies currently up to speed will have inevitably gained much ground before this volume makes it to press. Therefore, we have put together a repository where we store our recent educational materials and do-it-yourself tutorials for users who wish to implement and publish models of Linguistic Linked Open Data with their own resources.[47]

Summary

This chapter provides a general introduction to Linked Data and its application in the language sciences, with a specific emphasis on its uses for studying under-resourced languages. We identified characteristics of data for such languages, focusing on lexical resources (wordlists and dictionaries) and on annotated corpora (glosses and corpora). We further discussed aspects of resource integration, before focusing on Linked Data and under-resourced language data in particular. We then homed in on Linked Data for linguistics and NLP, and we gave two brief case studies of linguistic data sources that have been transformed into Linked Data. Finally, we described in detail the status and the bandwidth of applications of Linked Open Data technologies to under-resourced languages in the general context of the Open Linguistics Working Group and the developing Linguistic Linked Open Data (LLOD) ecosystem.

Notes

1. http://tei-c.org.

2. https://www.iso.org/developing-standards.html.

3. http://linguistics.okfn.org/.

4. https://www.w3.org/community/ontolex/.

5. http://www.lider-project.eu/.

6. Furthermore, increased access to language descriptions leads to increased documented typological diversity (at least in phonology, cf. Moran 2012a).

7. http://www.endangeredlanguages.com.

8. Important language catalogs include the Ethnologue (Lewis, Simons, and Fennig 2014) and the Open Languages Archive Network (OLAC).

9. http://glottolog.org.

10. Any concrete definition of the "under-resourced-ness" of languages' data should probably include a checklist of data types, as in "language X has a grammar, a dictionary, a corpus, a treebank." This definition would be problematic because what we know about worldwide language documentation is dynamic. Not only is documentation increasing, it is also decreasing, as for instance when the last records of language X are encoded in no longer accessible (electronic) formats.

11. Even more frightening for linguists studying linguistic diversity is that around one-third of the currently spoken languages are believed to be language isolates, or languages that are the last remaining leaf node in their language family tree. When lost, these languages take with them any typological structures that may not be accounted for anywhere else in the world. This phenomenon has often been compared to the loss of a biological species, which thereby limits biologists' view and study of the evolutionary processes that lead to worldwide diversity.

12. http://www.ruscorpora.ru/en/.

13. Basque, Bulgarian, Catalan, Croatian, Danish, Estonian, Finnish, Galician, Greek, Norwegian, Portuguese, Romanian, Serbian, Slovak, Slovene.

14. Icelandic, Irish, Latvian, Lithuanian, Maltese.

15. QuantHistLing is a project that has extracted wordlist data from many resources and uses both Linked Data and the ontological model called Lexicon Model for Ontologies (*lemon*; McCrae et al. 2010, 2011) (http://lemon-model.net/) to combine data sources.

16. http://corpora.informatik.uni-leipzig.de/.

17. Numerous examples: http://odin.linguistlist.org.

18. http://www-01.sil.org/computing/toolbox/.

19. http://dingo.sbs.arizona.edu/~sandiway/treebanksearch/.

20. http://www.mlta.uzh.ch/en/Projekte/Baumbanken.html.

21. The term "resource" is ambiguous: *Linguistic* resources are structured collections of data that can be represented, for example, in RDF. In RDF, however, "resource" is the conventional name of a node

in the graph, because, historically, these nodes were meant to represent objects described by metadata. In ambiguous cases, we use the terms "node" or "concept" whenever *RDF* resources are meant.

22. One example is the General Ontology of Linguistic Description (GOLD) by Farrar and Langendoen (2003).

23. For example, structured output from frameworks like Head-driven Phrase Structure Grammar (HPSG) or Lexical Functional Grammar (LFG).

24. http://clld.org.

25. http://www.lexvo.org/.

26. http://glottolog.org.

27. http://phoible.org.

28. http://panlex.org/.

29. http://www.acoli.informatik.uni-frankfurt.de/liodi/.

30. QuantHistLing was funded from 2010 to 2014 by the European Research Council (Michael Cysouw, University of Marburg, primary investigator). Its aims were to uncover and clarify phylogenetic relationships between native South American languages, particularly the Tukonoan, Witotoan, and Jivoroan language families, using quantitative methods. The two main objectives were the digitalization of the lexical resources on native South American languages and the development of innovative computer-assisted methods to quantitatively analyze this information. The project focused on formalizing (i.e., computationally coding) aspects both of data transformation and of the comparative method, by collaborating with research scientists in other fields.

31. Data are online at http://cysouw.de/home/quanthistling.html.

32. For a broad application of a translation graph aimed at worldwide coverage, see PanLex (Kamholz, Pool, and Colowick 2014): http://panlex.org.

33. Another necessary step is the identification of cognates via shared sound correspondences—a signal of genealogical relatedness. This process is comparable to DNA string comparison algorithms from bioinformatics, which have been reapplied and recoded for linguistic purposes (cf. List and Moran 2013).

34. There is an endpoint at http://www.linked-data.org:8890/sparql.

35. QuantHistLing data available in RDF and *lemon*: http://www.linked-data.org/datasets/qhl.ttl.zip.

36. http://phoible.org.

37. http://clld.org/datasets.html.

38. CLLD applications can conveniently use the Github "pull" functionality; in other words, CLLD project-specific applications can retrieve data directly from online hosted data and code repositories.

39. https://github.com/RDFLib/rdflib.

40. http://nbviewer.ipython.org/gist/xflr6/9050337/glottolog.ipynb.

41. There are several RDF serialization formats (e.g., Turtle, N-triples, XML). We do not go into detail with regard to them here.

42. Full SPARQL functionality is not supported. See: http://portal.clld.org/.

43. See Moran and Cysouw (2018) for a systematic exposition.

44. Originally, LLOD metadata was maintained under http://datahub.io. At the time of writing, LLOD metadata is being maintained under http://linghub.org. Because the LLOD cloud diagram is

now generated as a view of the LOD cloud diagram, novel datasets can be added via https://lod-cloud.net/add-dataset.

45. For example, complex copyright situations may arise if one resource (say, a lexicon) were to be developed on the basis of a second resource (say, a newspaper archive) and researchers felt uncertain whether the examples from the original newspaper contained in the lexicon violate the original copyright. Ethical problems may arise if a database of quotations from a newspaper were linked to a database of speakers and that database were further connected with, say, obituaries from the same newspaper. Even if this were done only in order to study generation-specific language variation, one may wonder whether such an accumulation of information violates the privacy of the people involved.

46. https://languagelink.let.uu.nl/tds/.

47. http://acoli.informatik.uni-frankfurt.de/resources/llod/index.html.

References

Abromeit, F., C. Chiarcos, C. Fäth, and M. Ionov. 2016. "Linking the Tower of Babel: Modelling a Massive Set of Etymological Dictionaries as RDF." In *Proceedings of the 5th Workshop on Linked Data in Linguistics (LDL-2016): Managing, Building and Using Linked Language Resources*, 11–19. Portoroz, Slovenia, ELRA.

Benson, M. 1990. "Collocations and General-Purpose Dictionaries." *International Journal of Lexicography* 3 (1): 23–34.

Bickel, B., and J. Nichols. 2015. Autotyp. http://www.autotyp.uzh.ch/.

Brown, C. H. 2013. "Hand and Arm." In *The World Atlas of Language Structures Online,* edited by M. S. Dryer and M. Haspelmath. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Carlson, L., M. E. Okurowski, D. Marcu, L. D. Consortium et al. 2002. RST Discourse Treebank. Linguistic Data Consortium, University of Pennsylvania.

Chiarcos, C. 2008. "An Ontology of Linguistic Annotations." *LDV Forum* 23 (1): 1–6.

Chiarcos, C., M. Ionov, M. Rind-Pawlowski, C. Fäth, J. W. Schreur, and I. Nevskaya. 2017. "LLOD-ifying Linguistic Glosses." In *International Conference on Language, Data and Knowledge*, 89–103. Galway, Ireland. Springer: Cham.

Chiarcos, C., J. McCrae, P. Cimiano, and C. Fellbaum. 2013. "Towards Open Data for Linguistics: Linguistic Linked Data." In *New Trends of Research in Ontologies and Lexical Resources*, edited by A. Oltramari, Lu-Qin, P. Vossen, and E. Hovy. Heidelberg: Springer.

Chiarcos, C., S. Nordhoff, and S. Hellmann. 2012. *Linked Data in Linguistics*. Berlin, Heidelberg: Springer.

Clark, A. 2003. "Combining Distributional and Morphological Information for Part of Speech Induction." In *Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics*, 59–66. Association for Computational Linguistics.

Cysouw, M. 2005. "Quantitative Methods in Typology." In *Quantitative Linguistics: An International Handbook*, edited by G. Altmann, R. Köhler, and R. G. Piotrowski, 554–578. Berlin: Walter de Gruyter.

de Melo, G. 2015. "Lexvo.org: Language-Related Information for the Linguistic Linked Data Cloud." *Semantic Web Journal* 6 (4): 393–400.

Dimitriadis, A., M. Windhouwer, A. Saulwick, R. Goedemans, and T. Bíró. 2009. *How to Integrate Databases without Starting a Typology War: The Typological Database System. The Use of Databases in Cross-Linguistic Studies*, 155–207. Berlin: Mouton de Gruyter.

Donohue, M., R. Hetherington, J. McElvenny, and V. Dawson. 2013. World phonotactics database. Department of Linguistics, Australian National University. http://phonotactics.anu.edu.au.

Dryer, M. S., and M. Haspelmath. 2013. *WALS Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Evans, N., and S. C. Levinson. 2009. "The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science." *Behavioral and Brain Sciences* 32:429–448.

Farrar, S., and T. Langendoen. 2003. "A Linguistic Ontology for the Semantic Web." *GLOT* 7 (3): 97–100.

Forkel, R. 2014. "The Cross-Linguistic Linked Data Project." In *Proceedings of the Third Workshop on Linked Data in Linguistics (LDL 2014)*, 60–66. Reykjavik, Iceland, ELRA.

Gangemi, A., R. Navigli, and P. Velardi. 2003. "The OntoWordNet Project: Extension and Axiomatization of Conceptual Relations in WordNet." In *Proceedings of On the Move to Meaningful Internet Systems (OTM2003)*, edited by R. Meersman and Z. Tari, 820–838. Catania, Italy.

Good, J. 2013. "Fine-Grained Typological Investigation of Grammatical Constructions Using Linked Data." In *Proceedings of the Tenth Biennial Conference of the Association of Linguistic Typology (ALT X),* Leipzig.

Hammarström, H., R. Forkel, M. Haspelmath, and S. Bank. 2015. Glottolog 2.6. Jena: Max Planck Institute for the Science of Human History. http://glottolog.org.

Ide, N., and J. Pustejovsky. 2010. "What Does Interoperability Mean, Anyway? Toward an Operational Definition of Interoperability." In *Proceedings of the Second International Conference on Global Interoperability for Language Resources (ICGL 2010)*, Hong Kong.

Ide, N., and K. Suderman. 2007. "GrAF: A Graph-Based Format for Linguistic Annotations." In *Proceedings of the 1st Linguistic Annotation Workshop (LAW 2007)*, Prague, Czech Republic. Association of Computational Linguistics.

ISO 639-3:2007. Codes for the representation of names of languages—Part 3: Alpha-3 code for comprehensive coverage of languages. Geneva: International Organization for Standardization.

Jones, D., and R. Havrilla. 1998. "Twisted Pair Grammar: Support for Rapid Development of Machine Translation for Low Density Languages." In *Machine Translation and the Information Soup*, edited by D. Farwell, E. Hovy, and L. Gerber, 318–332. Berlin: Springer.

Kamholz, D., J. Pool, and S. M. Colowick. 2014. "PanLex: Building a Resource for Panlingual Lexical Translation." In *Proceedings of the Ninth Language Resources and Evaluation Conference (LREC 2014)*, 3145–3150. Reykjavik, Iceland, ELRA.

Kingsbury, P., and M. Palmer. 2002. "From TreeBank to PropBank." In *Proceedings of the Third Language Resources and Evaluation Conference (LREC 2002)*, 1989–1993. Las Palmas de Gran Canaria, Canary Islands, Spain ELRA.

Kučera, H., and W. N. Francis. 1967. *Computational Analysis of Present-day American English*. Providence, RI: Brown University Press.

Lewis, M. P., G. F. Simons, and C. D. Fennig. 2014. *Ethnologue: Languages of the World, Seventeenth edition*. Dallas: SIL International.

List, J.-M., and S. Moran. 2013. “An Open Source Toolkit for Quantitative Historical linguistics.” In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, (ACL 2013), 13–18. Sofia, Bulgaria, Association of Computational Linguistics.

Matisoff, J. A. 2015. Sino-tibetan etymological dictionary and thesaurus (stedt). http://stedt.berkeley.edu/.

Maxwell, M., and B. Hughes. 2006. “Frontiers in Linguistic Annotation for Lower-Density Languages.” In *Proceedings of the Workshop on Frontiers in Linguistically Annotated Corpora 2006*, 29–37. Sydney, Australia,.Association of Computational Linguistics.

McCrae, J., G. Aguado-de Cea, P. Buitelaar, P. Cimiano, T. Declerck, A. G. Pérez, J. Gracia, et al. 2010. *The Lemon Cookbook*. Technical report, CITEC, Universität Bielefeld, Germany.

McCrae, J., D. Spohr, and P. Cimiano. 2011. “Linking Lexical Resources and Ontologies on the Semantic Web with Lemon.” In *The Semantic Web: Research and Applications*, *Proceedings of the 2nd European Semantic Web Conference (LNCS 3532)*, 245–259. Springer.

McNew, G., C. Derungs, and S. Moran. 2018. “Towards Faithfully Visualizing Global Linguistic Diversity.” In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 805–809. May 7–12, Miyazaki, Japan. http://www.lrec-conf.org/proceedings/lrec2018/pdf/813.pdf.

Meyers, A., N. Ide, L. Denoyer, and Y. Shinyama. 2007. “The Shared Corpora Working Group Report.” In *Proceedings of the First Linguistic Annotation Workshop (LAW-I), held in conjunction with ACL-2007,* 184–190. Prague, Czech Republic. Association of Computational Linguistics.

Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. “Annotating Noun Argument Structure for NomBank.” In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, 803–806. Lisbon, Portugal, ELRA.

Moran, S. 2009. “An Ontology for Accessing Transcription Systems (OATS).” In *Proceedings of the First Workshop on Language Technologies for African Languages (AfLaT 2009)*, Athens, Greece. Association for Computational Linguistics.

Moran, S. 2012a. “Phonetics Information Base and Lexicon.” PhD diss., University of Washington.

Moran, S. 2012b. “Using Linked Data to Create a Typological Knowledge Base.” In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann,129–138. Berlin: Springer.

Moran, S., and M. Brümmer. 2013. “Lemon-Aid: Using Lemon to Aid Quantitative Historical Linguistic Analysis.” In *Proceedings of the Second Workshop on Linked Data in Linguistics: Representing and Linking Lexicons, Terminologies and Other Language Data*, 28–33. Pisa, Italy, Association of Computational Linguistics.

Moran, S., and M. Cysouw. 2018. “The Unicode Cookbook for Linguists: Managing Writing Systems Using Orthography Profiles.” *Translation and Multilingual Natural Language Processing series in Language Science Press.* DOI: https://doi.org/10.5281/zenodo.1296780; http://langsci-press.org/catalog/book/176.

Moran, S., D. McCloy, and R. Wright. 2014. *PHOIBLE Online*. Leipzig: Max Planck Institute for Evolutionary Anthropology.

Nordhoff, S., H. Hammarström, R. Forkel, and M. H., eds. 2013. Glottolog 2.2. Leipzig: Max Planck Institute for Evolutionary Anthropology. http://glottolog.org.

Pederson, T. 2008. “Empiricism Is Not a Matter of Faith.” *Computational Linguistics* 34 (3): 465–470.

Poornima, S., and Good, J. 2010. “Modeling and Encoding Traditional Wordlists for Machine Applications.” In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, 1–9. Uppsala, Sweden, Association for Computational Linguistics.

Pradhan, S. S., L. Ramshaw, R. Weischedel, J. MacBride, and L. Micciulla. 2007. “Unrestricted Coreference: Identifying Entities and Events in OntoNotes.” 1st *IEEE International Conference on Semantic Computing (ICSC)*, 446–453. Irvine, CA, IEEE.

Prasad, R., N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. K. Joshi, and B. L. Webber. 2008. “The Penn Discourse TreeBank 2.0.” In *Proceedings of the Sixth Language Resource and Evaluation Conference (LREC 2008)*, 2961–2968. Marrakesh, Morocco.

Prud’hommeaux, E., and A. Seaborne. 2008. SPARQL Query Language for RDF. W3C Recommendation January 15, 2008.

Pustejovsky, J., P. Hanks, R. Sauri, A. See R. Gaizauskas, A. Setzer, et al. 2003. “The TimeBank Corpus.” In *Proceedings of Corpus Linguistics 2003. UCREL technical paper number 16*, 647–656. UCREL, Lancaster University, UK.

Rehm, G., and H. Uszkoreit. 2013. *META-NET Strategic Research Agenda for Multilingual Europe 2020*. Berlin: Springer.

Schalley, A. C. 2012. “TYTO—A Collaborative Research Tool for Linked Linguistic Data.” In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 139–149. Berlin: Springer.

Snyder, B., R. Barzilay, and K. Knight. 2010. “A Statistical Model for Lost Language Decipherment.” In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1048–1057. Uppsala, Sweden, Association for Computational Linguistics.

Taylor, A., M. Marcus, and B. Santorini. 2003. “The Penn Treebank: An Overview.” In *Treebanks (Text, Speech and Language Technology)*, edited by A. Abeillé, vol. 20, 5–22. Dordrecht: Springer.

Tiedemann, J. 2012. “Character-Based Pivot Translation for Under-Resourced Languages and Domains.” In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 141–151 (EACL 2012). Avignon, France: Association for Computational Linguistics.

Windhouwer, M., and S. Wright. 2012. “Linking to Linguistic Data Categories in ISOcat.” In *Linked Data in Linguistics*, edited by C. Chiarcos, S. Nordhoff, and S. Hellmann, 99–107. Berlin: Springer.

Wright, S. 2004. “A Global Data Category Registry for Interoperable Language Resources.” In *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC 2004)*, 123–126. Lisboa, Portugal.

Yarowsky, D., G. Ngai, and R. Wicentowski. 2001. “Inducing Multilingual Text Analysis Tools via Robust Projection across Aligned Corpora.” In *Proceedings of the First International Conference on Human Language Technology Research*, 1–8. San Diego, CA, Association of Computational Linguistics.

Zock, M., and S. Bilac. 2004. “Word Lookup on the Basis of Associations: From an Idea to a Roadmap.” In *COLING 2004: Enhancing and Using Electronic Dictionaries*, edited by M. Zock, 29–35. Geneva, Switzerland: Association of Computational Linguistics.